The “Timeline” Method of Studying Electoral Dynamics

by

Christopher Wlezien, Will Jennings, and Robert S. Erikson

1. Author affiliation information

CHRISTOPHER WLEZIEN is Hogg Professor of Government at the University of Texas at Austin, USA

WILL JENNINGS is Professor of Political Science and Public Policy, University of Southampton, UK.

ROBERT S. ERIKSON is Professor of Political Science at Columbia University. New York, USA

2. Acknowledgements

Previously presented at the conference on “Innovations in Comparative Political Methodology” at Texas A&M University, College Station, March 25-27, 2015. Thanks to Jude Hays, Simon Jackman, Gary King, Michael Lewis-Beck, Mark Pickup, Vera Troeger, and Guy Whitten for comments on earlier versions of this paper.

3. Corresponding author contact information

*Address correspondence to Christopher Wlezien, Department of Government, University of Texas at Austin, Austin, TX 78712-1704, USA. E-mail: Wlezien@austin.utexas.edu.

4. Article Title

The “Timeline” Method of Studying Electoral Dynamics

5. Short title

The “Timeline” Method

6. Keywords

Polls; votes; predictability; time horizons; forecasts.

Abstract

To study the evolution of electoral preferences, Erikson and Wlezien (2012) propose assessing the correspondence between pre-election polls and the vote in a set of elections. That is, they treat poll data not as a set of time series but as a series of cross-sections—across elections—for each day of the election cycle. This "timeline" method does not provide complete information, but does reveal general patterns of electoral dynamics, and has been applied to elections in numerous countries. The application of the method involves a number of decisions that have not been explicitly addressed in previous research, however. There are three primary issues: (1) how best to assess the evolution of preferences; (2) how to deal with missing data; and (3) the consequences of sampling error. This paper considers each of these issues and provides answers. In the end, the analyses suggest that simpler approaches are better. It also may be that a more general strategy is possible, in which scholars could explicitly model the variation in poll-vote error across countries, elections, parties and time. We consider that direction for future research in the concluding section.

How do voters' preferences evolve over the electoral cycle? Do preferences change? Do the changes last? The answers to these questions are important, as they could reveal how election outcomes come into focus. Indeed, they can shed light on whether and how election campaigns matter.

With data on the electorates' trial heat preferences over time, one's first thought might be to conduct time series analysis of pre-election polls of vote intentions. That is, we could examine the relationship between polls at different points in time within the various election years taken separately or pooled together. In theory, this would tell us much of what we want to know; in practice, it is not so straightforward because of data limitations.

There are two main reasons why studying electoral preferences as a time series is impractical. First, pre-election poll observations are missing for many days and even for weeks at a time. This has fairly obvious implications for what we can do with standard time series techniques. One cannot readily model time-series with large amounts of missing data. Second, with data based on survey samples, the ratio of error variance to the variance of the time series is quite large. This has substantial, if less obvious, complications: the presence of sampling (and other survey) error makes it difficult to uncover the underlying time series process. This is not to deny that a sequence of poll results can be treated as a statistical time series when polls are plentiful, as often is the case in the closing months of election cycles.

With statistical time series analysis often not feasible, what can we do instead? Wlezien and Erikson (2002) introduce a method that treats the poll data not as a set of time series but as a series of cross-sections—across elections—for each day of the election cycle. With the data organized as a series of cross sections, one can see how the vote across elections matches up with poll results at different points in the election cycle. This solution allows scholars to assess how informative polls throughout the election cycle are about the final vote, e.g., what polls 200 days before the election tell us, which is interesting unto itself. Scholars have used the method to study elections in the US (Erikson and Wlezien 2012), the UK (Wlezien et al. 2013) and in 45 different countries (Jennings and Wlezien 2016). They also have incorporated it into election forecasting models (Pickup and Johnston 2007; Armstrong et al. 2015; Graefe 2015; Rothschild 2015).[1] Its use may become even more prevalent as the number of elections and countries where poll data are available increases.

The application of this "timeline" method involves a number of decisions that have not been explicitly examined, however. Three primary issues require attention.

First, is there a preferred statistic for tracking and assessing the evolution of preferences over time? Past work has used different approaches. Wlezien and Erikson (2002) and Erikson and Wlezien (2012) rely on coefficients as well as the corresponding *R*-squareds from regressions relating the polls and the vote in their analysis of US presidential elections. Wlezien et al (2013) do much the same. Jennings and Wlezien (2016) focus on the regression root mean squared errors (RMSEs) in their comparative research. These statistics provide different information about the alignment of the polls and the final vote. They do not, however, reveal the closeness between the two.

Second, how should one deal with missing data in the time series? Different approaches have been used, some very basic and others much more difficult and demanding. We do not

[1] Also see Campbell (2008), Lewis-Beck and Stegmaier (2014), Lewis-Beck and Tien (2016).

know whether and how the difference matters. To what extent, if any, are findings distorted by interpolating missing data? Do fancier solutions like multiple imputation perform better?

Third, should one be greatly concerned about sampling error? And if so, how should one deal with it? As polling becomes denser over the election timeline, more polls and more respondents lessen the concern about measurement. But what about sampling error earlier in the timeline when polling data is thinner? And, how does the growing number of polls and respondents impact results over the timeline? The preceding research has not fully addressed this issue, at least not explicitly, and so we do not fully understand its impact.

This paper considers each of these issues and provides answers. In the end, the analyses suggest that simpler solutions are better. For assessing how polls predict elections, the mean absolute error between the vote and poll shares appears to be as informative as its regression-based alternatives, and it represents a more encompassing statistic. For dealing with missing data, basic linear interpolation works about as well as complicated and highly computationally-intensive alternatives, like multiple imputation. Finally, sampling error tends to be a fairly minor problem for the application of the timeline method given the variation in support we observe across parties, countries, and elections. This all is good news for research. It also may be that a more general modeling strategy is possible, one that would allow us to simultaneously explore differences across countries, parties and time itself, which we consider that direction for future research in the concluding section.

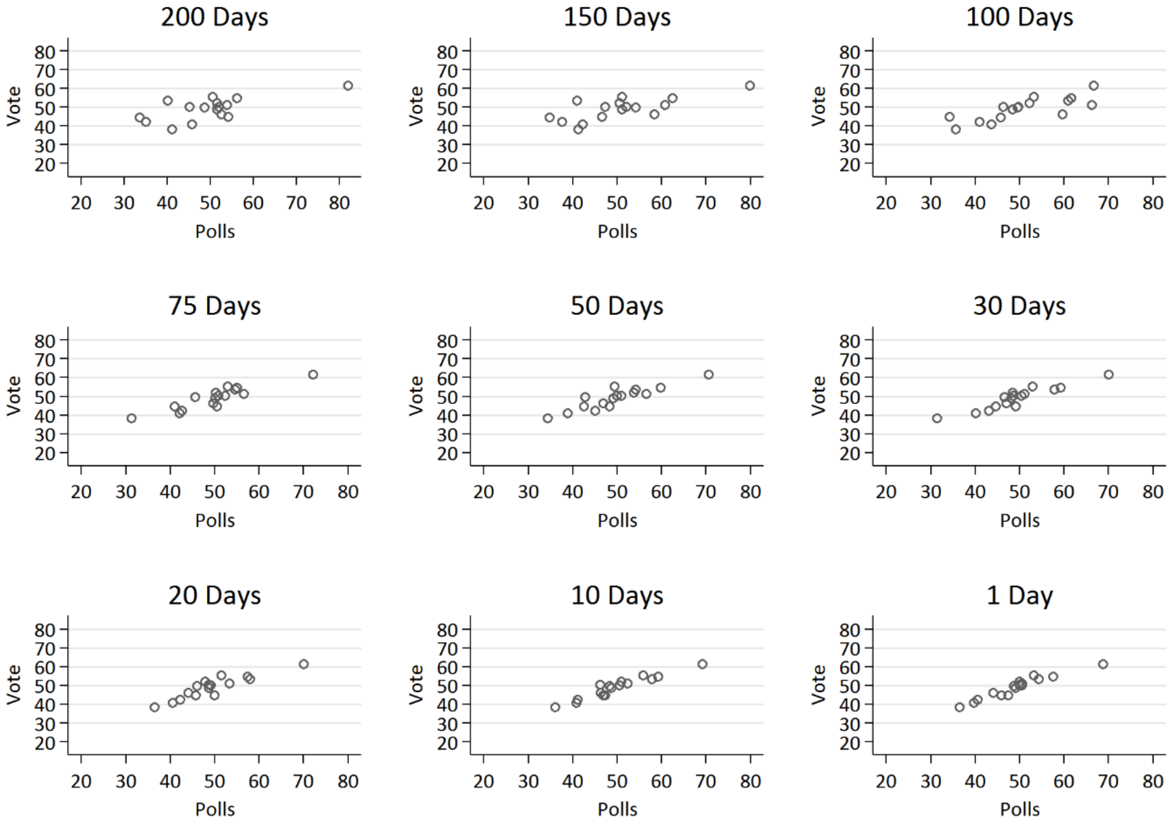
The Timeline Method

The Erikson-Wlezien method assesses the relationship between the vote in different elections and pre-election polls for each day of the election cycle, i.e., the "timeline" of elections. Various survey organizations regularly ask people about their vote intentions. The wording varies but respondents typically are asked "If there were a general election held tomorrow, which party (or candidate) would you vote for?" Typically the party (or candidate) names are listed. Questions differ in other ways, and this can matter, but the tricky bit for survey organizations is interviewing a representative sample of voters (e.g. AAPOR 2009; Sturgis et al. 2016).

We know that polls do pretty well at the end of the election cycle, but what about earlier? To assess the performance of polls over the course of the longer timeline of elections cycles, we can examine the match between the vote and the polls on a daily basis by pooling together data from different elections. To put it simply, we can compare the vote, say, for the set of US presidential elections – for which we have polls, of course – and poll results from the day before the election, two days before, three days before, and so on, as far back as we have poll data available. We then can see how the polls line up with the vote day by day.

Figure 1 plots vote shares by poll shares for Democratic Party candidates in all US presidential elections between 1952 and 2012. In the upper left-hand panel of the figure, using polls that are available 200 days before the election, more than six months before an election, we see that there already is a discernible pattern. That is, the poll shares and the vote shares are positively related, though there also is a good amount of variation. As we turn to polls later in the election cycle, moving horizontally and then vertically through the figure, a clearer pattern emerges; the poll share and final vote share line up. This is as one would expect if preferences change and a nontrivial portion lasts. But how much do preferences evolve?

Figure 1. Party Vote Share by Party Poll Share for Selected Days of the Election Cycle—US Presidential Elections, 1952-2012



Let us formally characterize the relationship between polls and the vote over the election timeline. In countries with two parties, scholars have relied on a simple bivariate equation relating one party's share of the two-party vote with the two-party vote in polls. Although more complex when there are multiple parties, the two-party template can be generalized for multiple parties or candidates. We can model the vote share for party or candidate *j* in election *k* in country *m* using vote intentions in the polls on each day of the timeline:

$$\boldsymbol{VOTE_{jkm} = a_{jmT} + b_T Poll_{jkmT} + \varepsilon_{jkmT},}$$

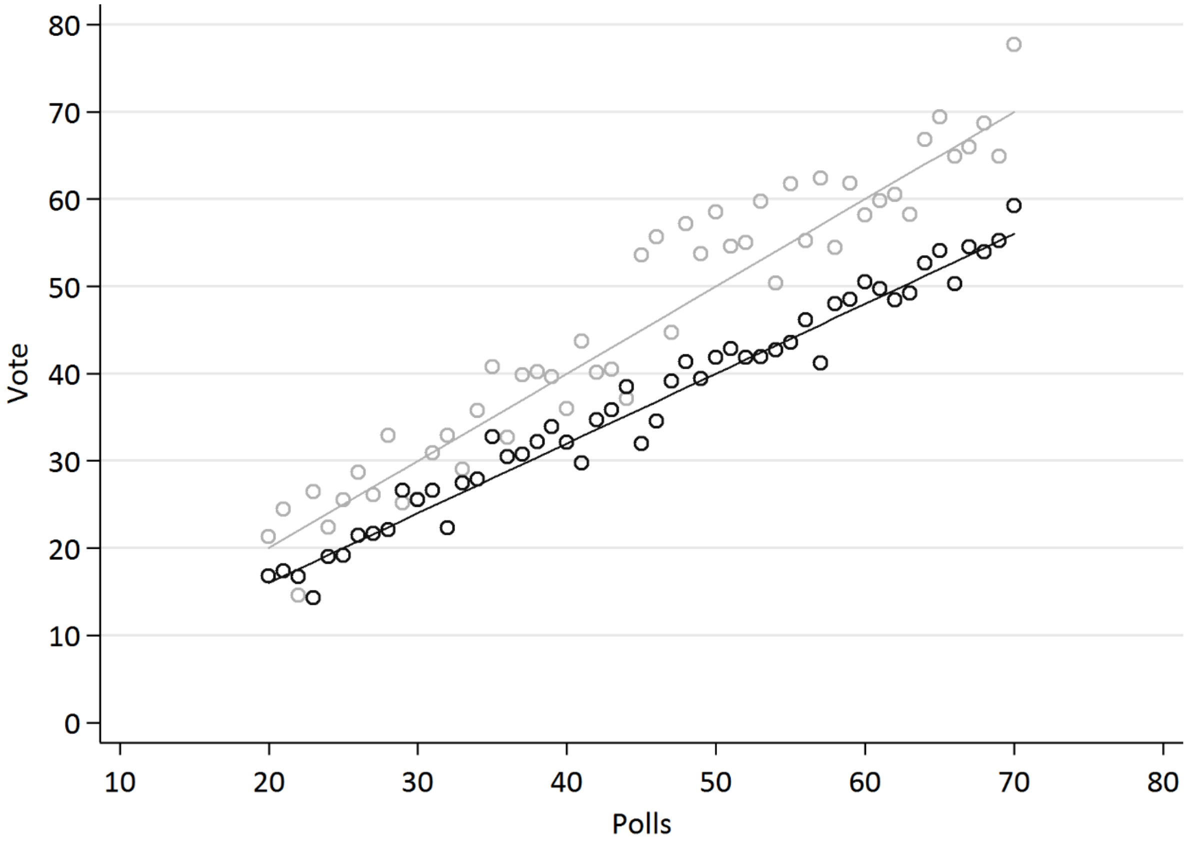
where *T* designates the number of days before Election Day and a_{jmT} represents a separate intercept for each party or candidate *j* in country *m*. This is important because the level of electoral support can vary systematically across parties. Let us assume that our timeline covers the year before Election Day. We would estimate an equation using polls from 365 days before each election, and then do the same using polls from 364 days in advance, and so on up to Election Day itself. Using the resulting estimates, we can see whether and how preferences come into focus over time. What statistic should we use to assess the match between polls and the vote?

Assessing Preference Evolution

In their analysis of US presidential elections, Wlezien and Erikson (2002) and Erikson and Wlezien (2012) rely on coefficients from regressions relating the polls and vote as well as the corresponding *R*-squareds to assess the evolution of electoral preferences. Wlezien et al (2013) do much the same in the examination of polls and the vote in the UK. By contrast, Jennings and Wlezien (2016) focus on the regression root mean squared errors (RMSEs) in their comparative research. As noted above, these provide different information about the alignment of the polls and the final vote. Let us consider these differences.

Clearly, the regression coefficient provides useful information. It tells us what proportion of the poll margin on each day carries forward to Election Day or, put differently, how much poll leads should be discounted. As the coefficient approaches 1.0 (and the intercept approaches 0), we expect that the poll margin provides an unbiased estimate of the final vote. Consider Figure 2, which depicts two sets of hypothetical observations, where both the polls and the vote differ. The slopes of the lines for the sets clearly differ, as the coefficient for one is 1.0 and the other is 0.8. For the former, indicated with light grey markers, we can "predict" the vote from raw poll results without making any adjustments. The polls are not perfect predictors, potentially because of survey error but also because preferences change in fairly random ways between the poll date and Election Day. For the second set of points in Figure 2, where b = 0.8, marked with black markers, the raw polls contain information about the vote but need to be adjusted in a systematic way. The coefficient implies that large poll leads decline by Election Day. Specifically, leads are expected to decline by 20%. For instance, a margin of 10 points in the polls should drop to 8 points when voters go to the ballot box.

Figure 2. Simulated data where b = 0.8 (σ = 2) and b = 1.0 (σ = 4)



Indicators of fit are useful too. Consider Figure 2 once again. We have seen that the two coefficients differ and this is important, but notice that the error variance also differs. That is, the residuals are larger where b = 1.0. This means that the vote is more *predictable* from the polls where *b* = 0.8. The *R*-squared provides one such measure of fit, and can be represented as one minus the sum of squared errors over the total sum of squares:

$$\boldsymbol{R^2 = 1 - \frac{SSE}{SST}}$$

In the case above, the *R*-squared is higher (0.97) where *b* = 0.8, with lower error variance, compared to that (0.93) for *b* = 1.0, with higher error variance. The relationship between polls and the vote is more deterministic in the former case – that is, electoral preferences change between the poll date and Election Day in a more predictable way.

The *R*-squared is a useful indicator of fit when comparing parties (or candidates) where vote shares are approximately the same on average, as the statistic is standardized to the total observed variance. For instance, it works well when studying the Democratic and Republican candidate vote in US presidential elections, per Erikson and Wlezien (2012). The *R*-squared is less useful when comparing parties in different countries, and especially across countries, where the variances in vote shares differ. Here, an unstandardized measure works better, specifically, the root mean squared error (RMSE). It is equal to the square root of the mean of the sum of squared errors:

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n} SSE}}$$

For the two sets of points depicted in Figure 2, the RMSE confirms what we see with the *R*-squared – when the latter is larger, the former is lower. (For the *b* = 0.8 line, the RMSE is 2.02, and for the *b* = 1.0 line, the number is 4.32.) This is not always true, of course.

The regression coefficients and measures of fit contain different information about the relationship between the polls and the vote, which we would like to encompass in a single measure.[2] The mean absolute error (MAE) offers a solution. It is the mean of the absolute error $|\boldsymbol{Poll_i - VOTE_i}|$ across *n* observations, where $\boldsymbol{Poll_i}$ is the poll share and $\boldsymbol{VOTE_i}$ is the vote share:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}|Poll_i - VOTE_i|}$$

The statistic directly captures the match between the polls and votes. It also has the advantage of being simple to calculate and easy to understand.[3] In Table 1 we report the MAE for the sets of data depicted in Figure 2, alongside the corresponding *R*-squared and RMSE. There we can see that the lowest MAE (3.62) is observed for the equation with *b* = 1.0, even though the measures of fit for that set of points indicate larger prediction errors. This is because the MAE taps the degree to which the final vote is evident from a naïve reading of the polls. It reveals how much aggregate preferences are crystallized. It also may be what we ultimately want to know.

It should be noted that in most uses of the timeline statistical apparatus, the observations are for different points in the timeline for a constant set of elections in a particular country. For these instances, the dependent variable (the vote) is a constant for all outcomes. This constrains the statistics to move together. If the regression coefficient increases from one point in the timeline to another, the *R*-squared will too, barring exceptional changes in the variance of the

[2] If we were solely interested in forecasting, by contrast, we would care more about fit than the regression coefficient *given the model*, as prediction errors would be critical.

[3] We also could employ the mean absolute squared error (MASE), which is a little more complicating and makes little difference in practice.

independent variable.[4] And, as the *R*-squared goes up, the RMSE must go down. The regression coefficient and *R*-squared, and the mean absolute error are less constrained to move together if the comparison is across sets of elections, where the variances of the dependent variable (vote) differ. Consider the polls-vote relationship in presidential elections by comparison with legislative ones, where outcomes differ. Here, we can observe a larger regression coefficient accompanied by a weaker fit, e.g., a lower RMSE, simply because of the differences in variances across the two sets of elections.

Table 1. Summary of *R*-squared, RMSE and MAE for Simulated Data

Equation	**Error distribution**	**n**	**Residual Sum of Squares**	**Total Sum of Squares**	**R-squared**	**RMSE**	**MAE**
Y = 0.8X + e	*μ=0, σ=2*	51	200.06	7357.24	0.97	2.02	8.55
Y = 1.0X + e	*μ=0, σ=4*	51	912.81	12592.37	0.93	4.32	3.62

[4] For instance, in analysis of the polls-vote relationship, it could be that the variance of the polls decreases over the timeline and the coefficient increases while the fit nevertheless decreases. This is not what we observe in practice (see Erikson and Wlezien 2002; Wlezien et al. 2013).

Dealing with Missing Data

As discussed, pre-election polls are sometimes sparse and conducted at irregular intervals.[5] What to do about missing data? There are different possible approaches. One option is to ignore the missing data and estimate the equation using available data. However, that makes the tracking of statistics over the timeline at the mercy of whichever polls are available for which days. Instead, Erikson and Wlezien (2012) use linear interpolation, as do Wlezien et al. (2013). Jennings and Wlezien (2016) use a method of linear interpolation plus random error, estimated using multiple imputation techniques, combined with bootstrapping.[6] There is a massive difference in difficulty between these methods, and the latter actually can be prohibitive for scholars without access to large-scale computing capacity.

Linear interpolation

When readings of electoral preferences are missing, we can interpolate daily voter preferences from available polls. For any date without a poll, an estimate is created as the weighted average from the most recent date of polling and the next date of polling. Weights are in proportion to the closeness of the surrounding earlier or later poll. This is the approach proposed by Erikson

[5] For the last 200 days in Jennings and Wlezien's (2016) data set, polls are missing on around 90% of days, and the percentage goes up as the timeline length is increased.

[6] It is worth noting that neither approach is helpful when forecasting, since logically one cannot actually forecast from values that are imputed using future observations. When forecasting, one can impute missing values by carrying forward previous observations. One also can impute based on time-serial, e.g., autoregressive, models. For a discussion, see Graefe et al (2014); also see Honaker and King (2010).

and Wlezien (2012). Specifically, given poll readings on days $\boldsymbol{t-\delta}$ and $\boldsymbol{t+\theta}$, the estimate for a particular day *t* is generated using the following formula:

$$\widehat{\boldsymbol{V}}_{\boldsymbol{t}} = \left\{\frac{[\boldsymbol{\delta} \times \boldsymbol{V}_{\boldsymbol{t-\delta}} + \boldsymbol{\theta} \times \boldsymbol{V}_{\boldsymbol{t+\theta}}]}{(\boldsymbol{\theta} + \boldsymbol{\delta})}\right\}$$

We thus are able include in our analysis any election cycle from the moment the first poll is conducted in that cycle. This would not be acceptable in conventional time series analysis, as interpolating would compromise the independence of observations. Given that the methodology is explicitly cross-sectional, there is no such problem—interpolating actually permits a more fine-grained analysis.

The main drawback of the approach is that we cannot assess whether dynamics differ across particular elections. This is by design: the approach treats the data as a set of cross sections, not time series per se, and so allows us to observe general patterns of evolution across elections. Importantly, it allows us to assess patterns of correspondence in different subsets of elections, e.g., across types of systems.[7]

[7] Some might think we should ignore aggregate poll results and limit our attention to individual-level data—the responses of survey responses in multiple polls. Multi-level survey analysis may be fine for some purposes but typically is infeasible, where individual survey responses are available for only a fraction of polls over the election timeline, and we rarely have self-reported vote choice against which to compare prior vote intention in fewer cases still, almost all clustered at the very end of the election cycle. As much as we would like to match individual-level

Multiple imputation

By contrast with simple interpolation, multiple imputation allows the incorporation of uncertainty. This is the approach of Jennings and Wlezien (2016). Here, a random component is introduced based on the poll variance to reflect uncertainty associated with the imputed values. In this case, the formula for interpolation is as follows:

$$\boldsymbol{\widehat{V}_t = \left\{\frac{[\delta \times V_{t-\delta} + \theta \times V_{t+\theta}]}{(\theta + \delta)}\right\} + \varepsilon}$$

where ε, is drawn from a defined distribution *N* (*μ, σ*^{2}*)*. Jennings and Wlezien (2016) estimate the underlying variance of all polls, once the country, party and election intercepts are controlled for (such that *μ=0, σ=3.394*).[8] But an alternative would be to specify this noise component in polling for a given party either due to its historical variance (i.e., within-country) or due to its variance within a given election cycle (i.e., within-country, within-election). Another approach would be to allow shocks to cumulate (i.e., for the data to follow an autoregressive process), to allow for drift in the polls the longer the gaps between poll observations.

preferences registered over the timeline with their vote choice later in numerous election years in numerous countries, it just is not possible given the existing data.

[8] Specifically, they estimate a regression of the poll share as a function of a separate intercept for each party or candidate *j* in election *k* in country *m*. The residuals of this equation provide our measure of underlying variance of the polls once the country-party-election equilibrium is taken into account:

$$\boldsymbol{Poll = a_{jkm} + \varepsilon.}$$

Single-imputation techniques, adding this noise component, still treat imputed values as known in the analysis. This would underestimate the variance of the estimates, thus overstating their precision (King et al. 2001). Multiple imputation (Rubin 1987) addresses this issue by averaging the coefficients across the imputed data series and adjusting the standard errors to reflect noise due to imputation and residual variance.[9]

Bootstrapping

When comparing timelines, for example across countries or across electoral systems or across different periods, we want to be confident that differences between them are significant. But standard procedures do not generate measures of uncertainty for the mean absolute error (MAE), *R*-squared or RMSE. Bootstrapping is a solution to this. It enables us to estimate the sampling distribution of our measures of goodness-of-fit. Given that our data on polling tends to be exhaustive, at least in most countries, it is reasonable to assume that our sample is representative of the population of polls (particularly from the period 200 days out from Election Day). Bootstrapping the estimates is thus quite straightforward, with the regression estimated for randomly drawn resamples (with replacement) of the data repeated *N* times for each day of the election cycle. From this we can observe the amount of uncertainty surrounding our estimates.

[9] Rubin (1987) shows that where γ is the rate of missing data, estimates based on m imputations have an efficiency that is approximately $(1 + \frac{\gamma}{m})^{-1}$. In our later analysis, polls are missing on around 90% of days, so we use 50 imputed data series, which implies a relative efficiency of 0.98 compared to an infinite number of imputations.

In computational terms, bootstrapping is a *highly* intensive approach. Where we are estimating the timeline equation over 200 days of the election cycle, 50 multiple imputations means that this has to be estimated 10,000 times. Where we also are bootstrapping the regression equation to estimate the standard errors of the *R*-squared or RMSEs – which enables us to determine whether the relationship between polls and the vote differs significantly across institutional settings – the total number of estimations increases based on the number of samples one draws. In our case, we draw 1,000 samples, which means a total of 10,000,000 regressions. In comparative analyses where absorbed regressions are used to control for a large number of party-country intercepts, the number of predictor variables is also large (up to around 200 parties for all elections in Jennings and Wlezien 2016). This approach thus is hyper-intensive in terms of processing.

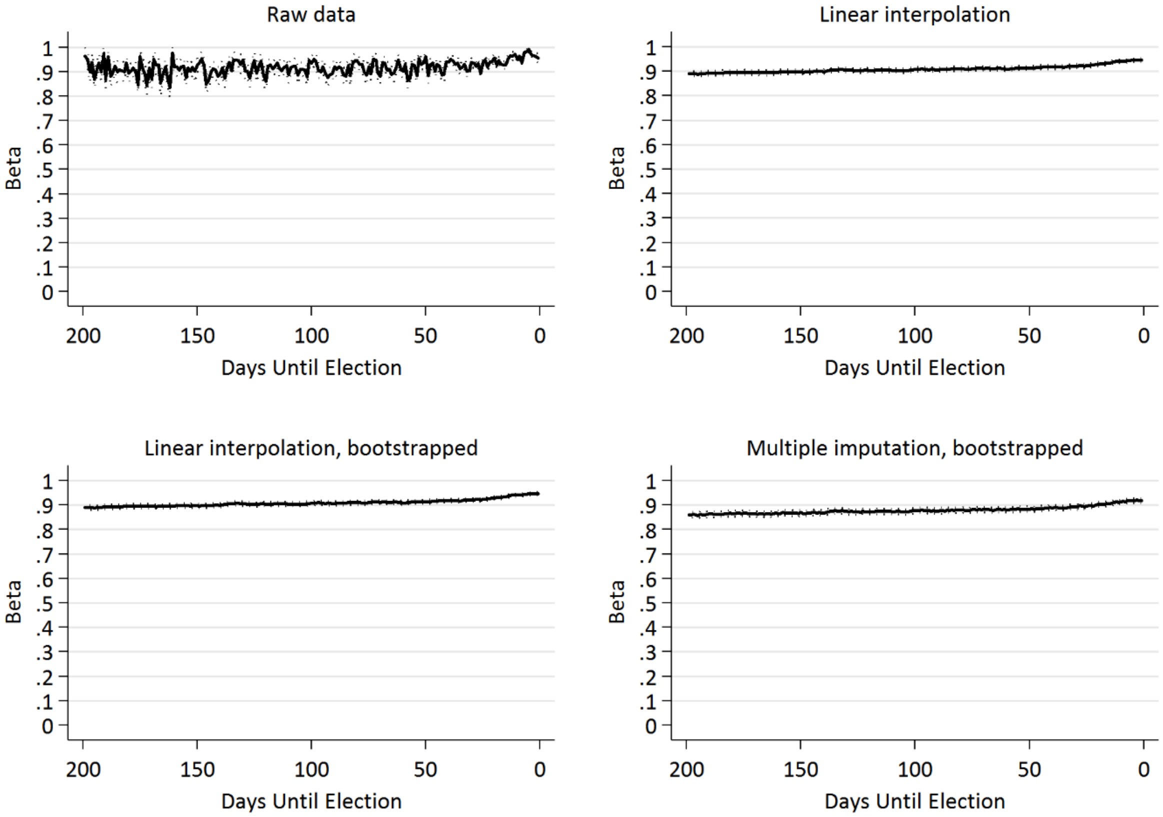
How do these approaches impact our analysis? This ultimately is an empirical question, of course, but do note that there are substantial differences in the average level of electoral support across parties and, to a lesser extent, elections. Consider analysis of variance (ANOVA) of available poll results, which indicate that 86% of the variance in party support in different countries and elections is due to systematic party differences. (See Appendix Table A1.) In actuality, since party variables are specific to each country, some of this party-explained variance reflects differences across countries; specifically, countries account for 44% and parties account for the other 42%. The remaining (14% of the total) variance reflects differences across elections and time, only a portion of which (10% of the total) is due to changes in preferences over the timeline. That is, most of the variance in the polls owes to differences across parties, countries and elections. Because of this, there is reason to think that the form of imputation makes little difference, i.e. adding in a little randomness probably will not make much difference to timeline analyses.

Comparing Methods for Dealing with Missing Data

To test the effect of these different methods of imputing missing data, we draw on comparative data on polls and the vote in presidential legislative elections in 45 countries. We focus on elections for which we have poll readings beginning 200 days before Election Day, so as to avoid change in estimates due to the addition of cases over the timeline. This leaves us with 249 discrete election cycles and 210 parties, where we exclude those parties whose vote share is less than 5 per cent. Further details of the data are available in Jennings and Wlezien (2016). In these cases polls are missing on 92% of days. To see how imputation methods matter, we estimate the timeline equation using each method described above: (1) raw data, (2) linear interpolation, (3) linear interpolation plus bootstrapping, and (4) multiple imputation plus bootstrapping. This enables us to compare the relative gains for analysis of using different techniques.

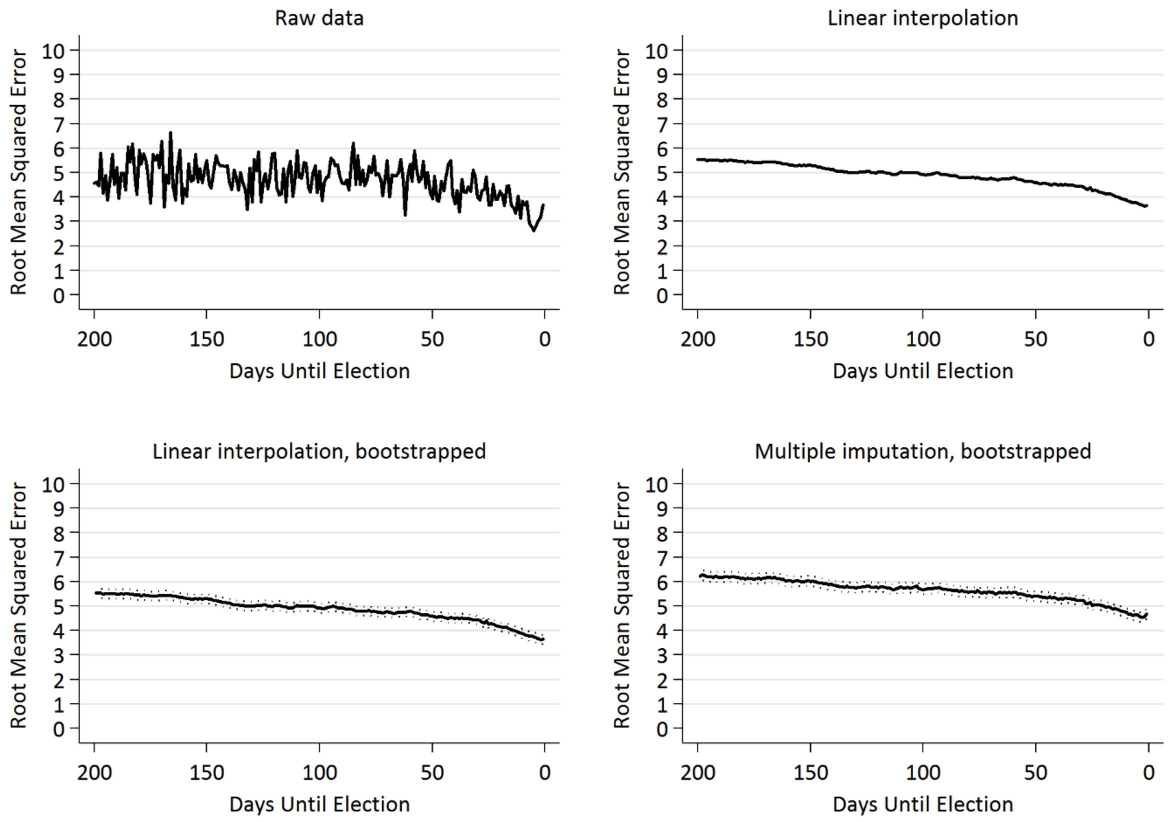
We first plot the regression coefficients estimated for the timeline equation using each of the methods. These are shown in Figure 3. Here we see that the essential patterns are the same. In the upper-left frame, we can see that estimates based on raw poll data, where the daily *N* is around 110 on average, are noisy, but still reveal the gradual increase in the *b* over the final two hundred days of the election timeline. The estimates are much smoother using the linear interpolation method, depicted in the upper-right frame, though they show essentially the same trend. The lower-left frame of Figure 3 shows the regression coefficient for linear interpolation plus bootstrapping. This step is not so critical for regression coefficients, for which standard errors are available, but is important for our estimates of RMSEs and MAEs, as we will see. (That is, it adds standard errors, allowing us to compare across time and also subsets of elections or parties.) The final, lower-right frame of the figure shows multiple imputation estimates, where the mean of the regression coefficient is slightly lower, due to the addition of noise to the underlying data, but reveals an identical pattern over time. Substantively, then, the linear interpolation step provides the largest gains in terms of smoothing the election timeline, overcoming the spottiness of data, with the multiple imputation step adding uncertainty about the true value of the coefficients.

Figure 3. Regression Coefficient (*b*) Predicting each Party's Vote Share from its Poll Share



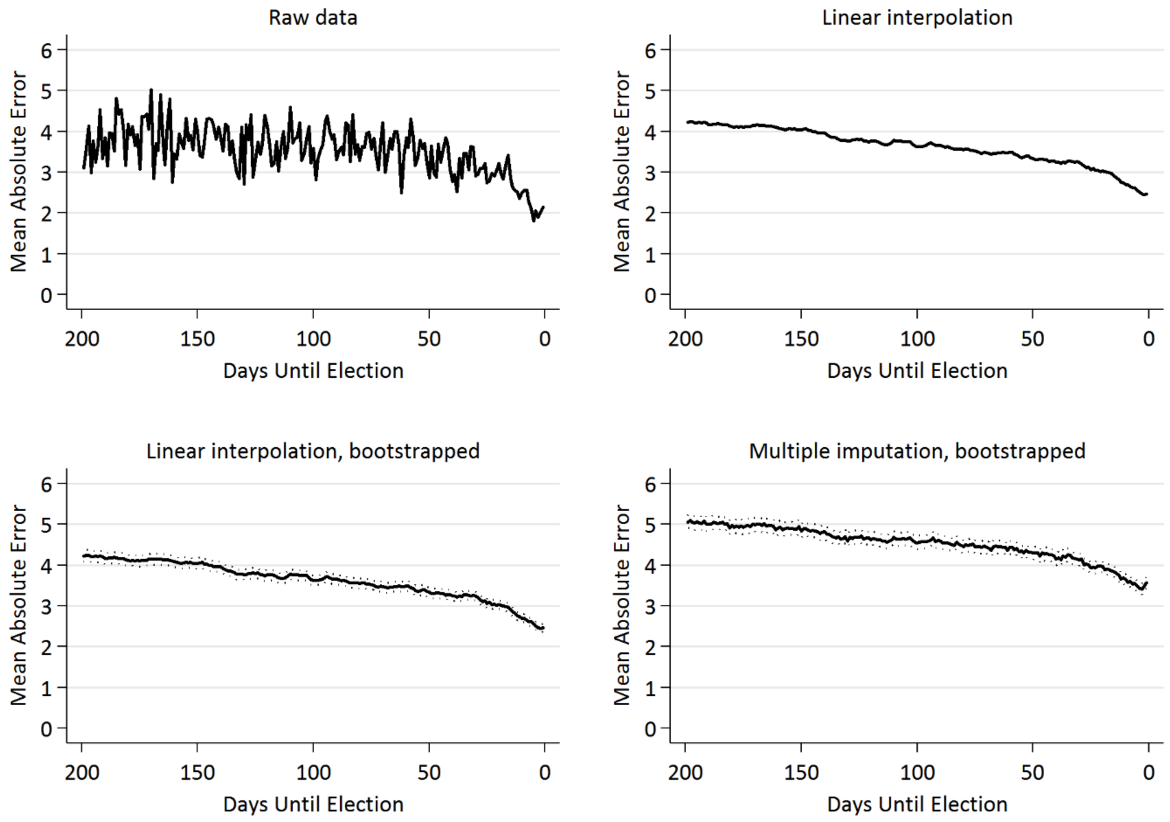
In Figure 4 we plot the RMSE for the timeline equation estimated using each of the different methods. Again, there is a good degree of commonality between the patterns revealed using each of the techniques. The trend is substantially noisier using raw data compared to any of the other methods, but reveals the same decline over time, accelerating over the final 30 days. Linear interpolation flattens the line to more clearly reveal the underlying trend and bootstrapping enables a direct assessment of the reliability of the estimates. The use of multiple imputation increases the level of the RMSE, due to its addition of uncertainty to the estimates, but the incline of the line is the same.

Figure 4. Root Mean Squared Errors for the Last 200 Days of the Election Cycle



Now, let us consider the estimated MAE using the different approaches. This is shown in Figure 5 for the same period. There we can see the now familiar pattern, where the estimates using the raw data bounce around and those using the three imputation approaches all look very similar, with the multiple imputation step adding error to the estimates but revealing essentially the same trend. It once again appears that what matters is that imputation is used, not the particular type of imputation one adopts.

Figure 5. Mean Absolute Errors for the Last 200 Days of the Election Cycle



Besides eyeballing the data, we can compare the distributions of the estimates, which are summarized in Table 2. The means confirm what we observed in the figures, specifically, that linear interpolation slightly decreases the regression coefficient and the estimated fit, i.e., the adjusted *R*-squared is lower and the RMSE higher than with the raw data, and multiple imputation widens this difference. Although the methods matter, the differences between them are not fundamental, as we saw in the figures. Indeed, the correlations between the betas, *R*-squareds, RMSEs and MAEs for each of the imputation-based methods are never less than 0.98, p<0.000 (see Appendix Table A2).[10]

Table 2. Summary Statistics of Estimates using Alternative Methods of Dealing with Missing Data, The Last 200 Days of the Election Cycle

	Betas	**Adjusted *R*-Squared**	**RMSE**	**MAE**
Raw data	0.919 (0.029)	0.917 (0.027)	4.673 (0.745)	3.528 (0.600)
Linear interpolation	0.909 (0.014)	0.911 (0.016)	4.877 (0.470)	3.631 (0.440)
Linear interpolation, bootstrapped	0.909 (0.014)	0.911 (0.016)	4.877 (0.470)	3.631 (0.440)
Multiple imputation, bootstrapped	0.879 (0.014)	0.881 (0.017)	5.647 (0.422)	4.530 (0.403)

Note: standard deviation in parentheses

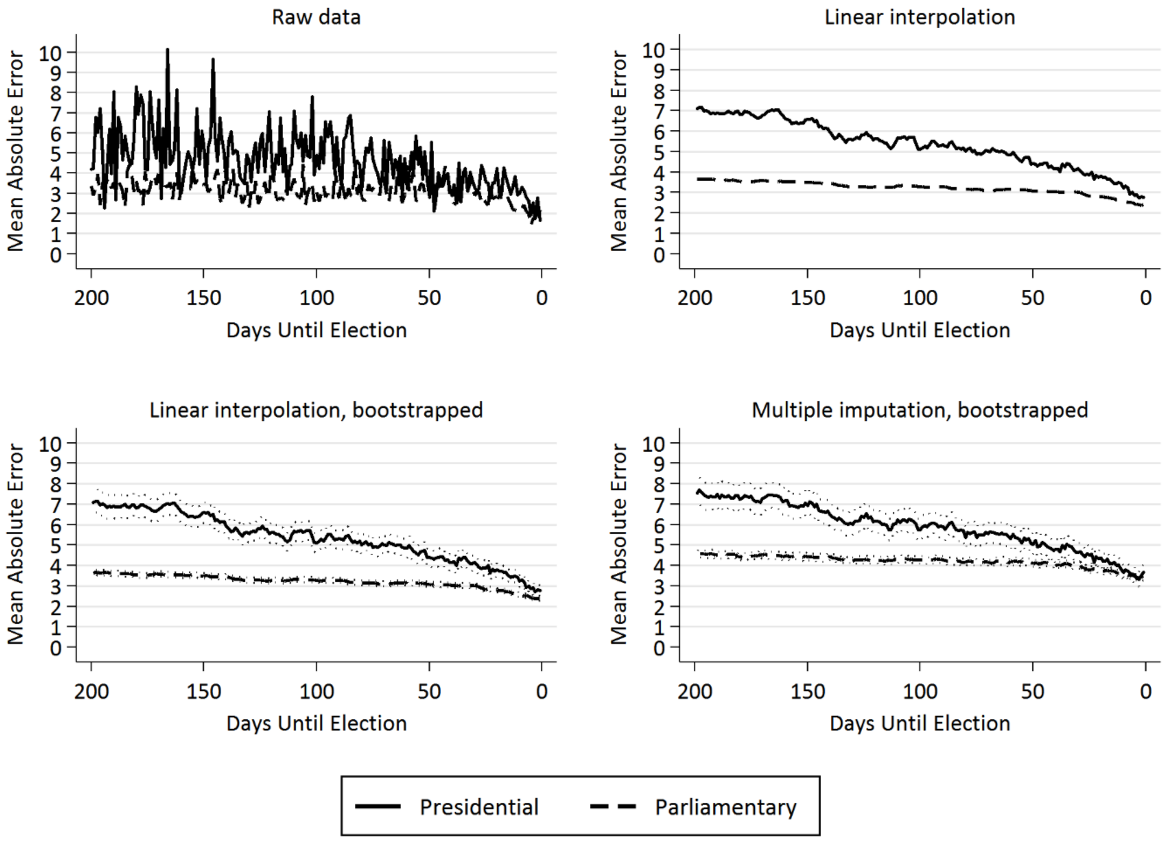
An Application

We also can assess how the different methods of dealing with missing data enable us to compare across particular types of election. In Figure 6 we plot the MAE using each of the methods for two subsets of elections – presidential and parliamentary. Disaggregating the data reduces the number of observations, which makes the estimates even noisier, especially when using raw poll

[10] That said, there are meaningful differences between estimates using raw data and those where missing values are imputed.

data. The interpolation step makes the difference between the timelines much clearer, whereby preferences are structured much earlier in parliamentary elections compared to presidential elections. The multiple imputation step makes little difference to the inferences that can be drawn about variation across types of election. Bootstrapping does, however, enable us to determine whether and at what point in the election cycle the differences are statistically significant. This is true both for linear interpolation and multiple imputation.

Figure 6. Mean Absolute Error for the Last 200 Days of the Election Cycle, Presidential vs. Parliamentary Elections



In summary, we have shown that imputation is consequential for understanding the vote-polls relationship. Firstly, it slightly dampens regression coefficients and increases unexplained variances, which more accurately depicts the true relationship between polls and the vote. The effect is largest in the step between linear interpolation and multiple imputation,
due to the addition of uncertainty/noise to the data. Imputation also decreases standard errors, however, and so it provides a clearer depiction of preference evolution and also enables slightly cleaner tests of differences across subsets of elections (or parties).

The benefits of more complex and highly computationally-intensive techniques like multiple imputation are less clear. That is, the same general inferences can be drawn using basic methods of linear interpolation. Admittedly, our example is a case where vote (and poll) shares are dominated by structural factors, i.e., parties, countries, and elections themselves. Under such circumstances, it does not matter much how one imputes. Where such differences are smaller, by contrast, the method may matter more. And when comparing subsets of elections, especially when the numbers of cases are smaller, the estimation of uncertainty can be important as well. This was the case for our application, after all. Our findings thus contribute to debates over the use of multiple imputation in political science (King et al. 2001; Lall 2016).

Adjusting for Measurement Error

Survey results are never exact. Even if polls are unbiased, they inevitably contain an amount of sampling error. Interviewing some finite number of voters produces estimates of aggregated voter preferences that represent the sum of the true vote division at the moment (among the population sampled) plus sampling error. The error diminishes the ability to predict the vote from the sample vote division for the specific point in the campaign timeline. The amount of sampling error is a direct function of the number of voters interviewed. When pooling several polls to create a meta-sample of several thousand respondents, the concern about sampling error is trivial. Concern rises with sample sizes as low as in the hundreds of cases.

For comparing polls across the campaign timeline, one can be concerned both about distortion from sampling error in general and also how the distortion might vary with the point in the timeline, since the density of surveys is greatest as Election Day looms. How much are the weaker estimates far in advance of election due to a sparse *N*?

If we were modeling campaign dynamics as a time series, the problem of error could be overwhelming, as each change in the polls is likely to contain far more error than truth. This is because true preferences evolve slowly over campaigns (Erikson and Wlezien 2012; Jennings and Wlezien 2016). In cross-sectional analysis, the range of true voter preferences for a point in the timeline (for different parties within different countries and/or different years) covers a wide range. Thus the ratio of error variance to true variance is less than with a time series approach. But should we still worry?

Fortunately, there is a way of estimating the sampling error of vote intentions for a party in a particular poll. To do so we assume the poll is conducted by simple random sampling. In practice, there are reasons why polls might both be less accurate and more accurate than by simple random sampling. Because polls are conducted by multi-stage random sampling, the sampling error by theory is a bit larger than when random sampling is assumed. On the other hand, pollsters can improve poll accuracy beyond what random sampling provides by post-stratifying the data, weighting respondents to maintain proportions of demographic groups that match the target population.[11] On balance, we assume that these competing forces cancel out. If we can adjust our statistical analysis for measurement error by assuming random sampling, we are better off than with no adjustment at all.

The assumption of simple random sampling means that every individual in the population has the same chance of being interviewed. Consider the poll of polls that generate the estimate of V for party j in election k in country m at time t. Let p = the proportion voting for party j, the party of interest and q = votes for all others. (Note, q = 1- p). From established statistical principles, the variance of the sampling error equals the observed within-sample variance divided by the number of cases, N. When the poll result is measured as proportion yes (p) or no (q) for a party or candidate, the formula for sample error variance reverts to $\frac{pq}{N}$. This quantity equals the error variance for the pooled measure of the aggregate proportion preferring party or candidate j at time t and election k and country m.

Next, we leverage the statistical fact that (assuming error variance is random), the total cross-sectional variance of V at time t equals the true variance of observed preferences V_t plus the error variance. The true variance can be backed out as the difference between the observed variance at t and the *average* error variance at t. The ratio of the true variance to the total variance equals the reliability of V_t.

[11] It also may be that pollsters “herd,” that is, adjusting their design or reporting practices in light of other pollsters’ results, which can reduce poll variation while increasing aggregate poll-vote errors (e.g., Silver 2014; Sturgis et al. 2016).

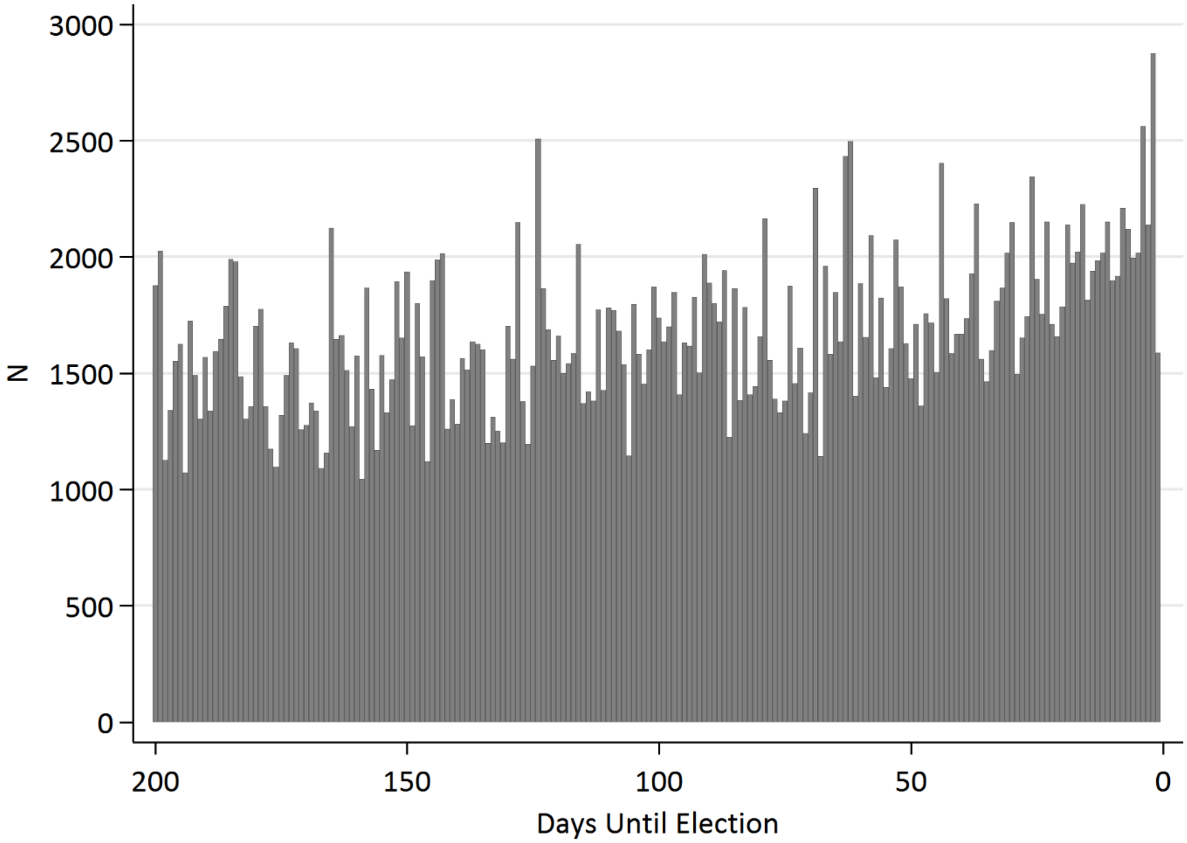
So far, we have described how sampling theory can be used to estimate the reliability of the cross-sectional preferences V_t. How does it affect our statistical instruments? For b, the coefficient from regressing election results on V_t, sampling error biases the estimated b downward by a factor of $\sqrt{reliability}$. The *R*-squared is biased downward in proportion to the *reliability* itself. For the observed difference (*G)* between the election result and the poll estimate, the expectation of the true differential is the square root of (*G* – error variance of V_t.). The mean *absolute* difference between the election result and poll estimate for the specific date in the timeline (mean absolute error) obviously is influenced by the mean error variance. Armed with estimates of survey error, one can estimate the reliability of our statistical instruments in general and for various points in the timeline. Where warned by low reliability estimates, we can adjust our statistical estimates accordingly. For example, STATA's program "eivreg" allows the researcher to estimate *b*'s, *R*-squareds, and RMSEs without bias from measurement error by plugging in the estimated reliabilities of the independent variables.[12]

[12] Reliability correction assumes that observations are unbiased, consisting of the intended scores, e.g., true voter preferences, plus measurement error. Errors for separate readings are assumed to be independent of each other. They also are assumed to be unrelated statistically to the intended target, here being true vote intentions. If this is not the case, as when a certain party is systematically given more support in polls than the facts warrant, this would be evident from the constant term in the equation predicting the vote but in any event would not affect any reliability calculation. Such bias could be apparent if the MAE statistic implies a degree of prediction that appears contrary to the other applied statistics.

The potential worry is that the "*N*"s of our pooled polls are low enough to bias statistical estimates, particularly the early dates in the timeline when the presence of multiple polls to pool is rare. Figure 7 shows the working *N* for all our observations over the 200-day timeline. This is the average pooled number of cases on each day in each country, based on actual poll data. Note that the working *N* per observation does increase over the timeline from about one thousand to fifteen hundred.[13]

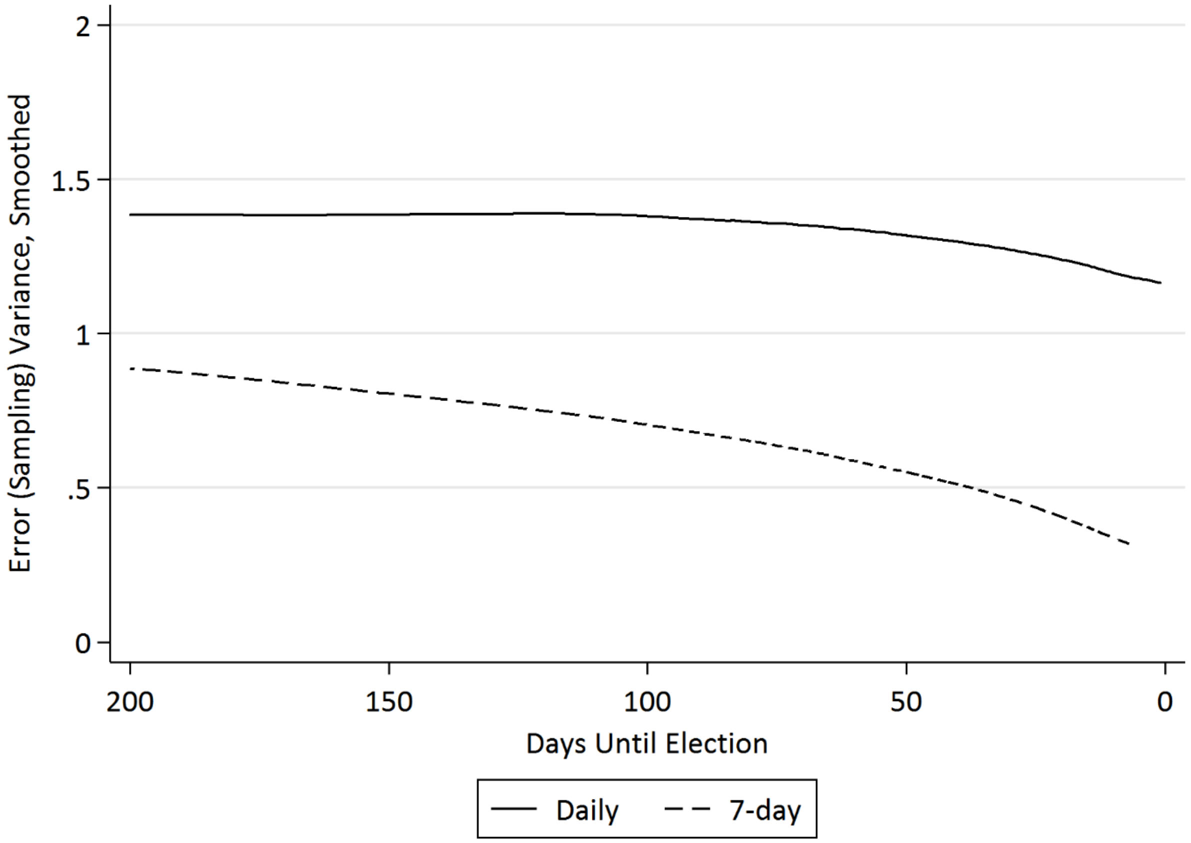
[13] The data are from Jennings and Wlezien (2016) and includes those polls – 5,000 polls in 16 of our 45 countries – where we have some information on the sample size of around, and treat missing values, which are approximately a quarter of the cases, as having an *N* of 1,000. Ten of the countries hold legislative elections – Australia, Austria, Croatia, Germany, Ireland, Japan, Norway, Sweden, UK, and the US – and seven hold presidential elections – Argentina, France, South Korea, Mexico, Philippines, Slovakia, and the US.

Figure 7. Daily Average of Poll *N*, All Elections (16 Countries)



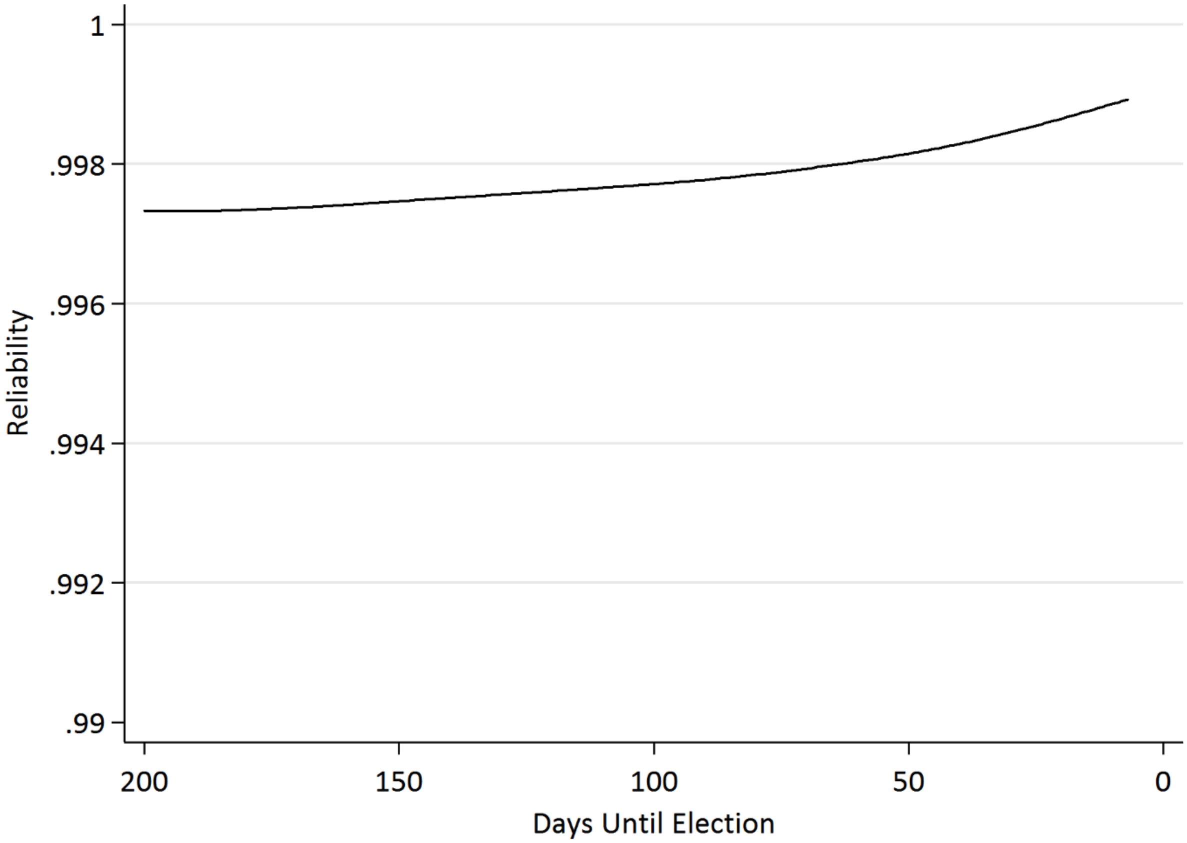
For any date in the timeline, one can estimate the error variance by assuming random assignment and the formula described above. Figure 8 presents the averages of these estimates of error variance per date in the timeline. The data are presented two ways—for polls centered on the date at hand, and for the pooled polls for the seven days ending on the indicated date. For daily readings, the average error variance is in the 1.0 to 1.5 range, meaning a confidence band of plus/minus 2 or 3 percentage points around the observed result. When measured for weekly data, the error variance drops below one point. How we adjust for reliability thus can make a difference. Importantly, though, by neither measure does the error variance rise precipitously for early dates in the timeline.

Figure 8. Daily and 7-Day Error Variance, LOWESS Curves, All Elections (16 Countries)



With the parties and candidates in our sample showing a wide range of vote percentages, the cross-sectional variance of the observed vote margins is very large, upwards of 100 percentage points. As a result, the reliability of the seven-day readings is absurdly high, well within the range from 0.99 to 1.00. See Figure 9. The good news from this is that the usual statistics predicting the vote from the sample—the regression coefficients, the *R*-squareds, and the RMSEs—need no correction. (Any estimate corrected for reliability would result in virtually no change.) And the reliability varies only trivially over the timeline and so has little effect on analysis of trends.[14]

Figure 9. Reliability of Vote Preferences, All Elections (16 Countries)



We must consider daily readings. The observed variance for a specific date is sketchy, due to the enormous missing data for the many dates when there is little or no polling. But as we see from Figure 8, the daily error variance is only slightly greater than when measured weekly.

[14] Reliabilities vary across parties according to their size, reflecting differences in error variance, which is predictably lower for small parties, and/or true variance, which also may be lower for small parties, at least where support is effectively bounded.

Offsetting the greater error variance when measured daily, the total variance of daily polls (when missing values are interpolated) also is larger. As a result, the reliability of daily readings should be about as high as when pooled for weekly readings. With error so small even for daily samples —and the large range of observed variance— the usual statistics need little or no correction for sampling error.

Discussion and Conclusion

Because of frequent missing data and an amount of sampling error that often dwarfs real change, the analysis of polls as a time series can be challenging. The "timeline" method offers a solution. By treating poll data not as a set of time series but as a series of cross-sections—across elections—for each day of the election cycle, researchers can observe how the vote matches up with poll results at different points in the election cycle. They thus can assess how preferences come into focus over time and also how informative polls throughout the election cycle are about the final vote. Indeed, as seen in the text, the method can be used to assess differences across types of elections, e.g., presidential vs. parliamentary.

Although the timeline method has been applied to good effect in previous research, we have seen that its application involves a number of decisions that are not directly addressed in that work. This paper concentrated on three of these: (1) the statistics for assessing preference evolution; (2) how to deal with missing data; and (3) the consequences of sampling error.[15]

[15] Although these issues have been considered separately here, it is possible to treat them simultaneously in a single statistical model (see Honaker and King 2010).

Previous research has relied on regression coefficients and related measures of fit, but we posit that the simple mean absolute error (MAE) has certain advantages over these, as it most directly captures the match between the polls across the campaign timeline and the final vote. That said, our analyses show that the various measures – betas, *R*-squareds, RMSEs and MAEs – are all highly correlated, revealing largely the same pattern in the evolution of preferences. This supports the use of basic approaches when characterizing the relationships between the polls and the vote.

For the treatment of missing data, there are different approaches. One can ignore the missing data and analyze what data are available. There also are various ways of imputing data, ranging from linear interpolation to multiple imputation. Our analyses consider the differences, and show that the step between raw data and linear interpolation is most crucial, as it more clearly reveals the underlying trend. Multiple imputation adds little to our analysis other than to increase the error attached to the estimates. Bootstrapping allows for comparisons of different subsets of parties, elections, time periods or other features of electoral choice (such as turnout or incumbent vs. opposition), but can be applied to any imputation technique, including linear interpolation.

Third, sample sizes differ over the election timeline and in systematic ways, i.e., the number (N) of respondents trends upward approaching Election Day. This is important because the N's can impact the match between polls and the vote independently of any underlying change in preferences, that is, the MAE will tend to increase as sampling error decreases. Our analysis provides a way of adjusting analysis to reflect polling intensity and also demonstrates that, while sampling error matters, it has minor impacts on demonstrated patterns of the poll-vote relationship over time.

Putting aside the issues addressed above, there is a limitation to the timeline method as employed to date. Using it to test hypotheses about differences across types of elections or political institutions or parties requires comparisons across subsets of elections, some of which are overlapping, e.g., where the proportionality of election systems is interrelated with the effective number of parties. This is difficult to examine by sub-setting cases, as one quickly loses statistical power. There is a more general modelling strategy, however. That is, one can treat the absolute poll-vote error as a dependent variable in a regression equation that estimates the simultaneous effects of various independent variables and time itself.[16] This parallels the sort of analysis presented in Figure 6 above, in demonstrating the different level and slope of the election timeline. What such an approach offers over previous analysis is the possibility of adding other variables to the model, and without reducing the number of observations as the subsets of cases become increasingly narrow. We thus can simultaneously assess differences across countries, parties and elections. For instance, we can test the effects of government and electoral institutions, characteristics of political parties, and over-time variation in electoral

[16] The equation might take the form: $\boldsymbol{|VOTE - POLL| = a + b_1T + b_2X + b_3X \times T + b_4Y + b_5Y \times T,}$ where the absolute error is a function of some intercept ($\boldsymbol{a}$) plus time ($\boldsymbol{T}$), i.e., the number of days before Election Day, plus some independent variable ($\boldsymbol{X}$) and its interaction with time ($\boldsymbol{X \times T}$) and another independent variable ($\boldsymbol{Y}$) and its interaction with time ($\boldsymbol{Y \times T}$).

context.[17]

It also is important to recognize that the polls-vote relationship over time is just one possible application of the timeline method. First, it can be used to with the same vote dependent variable but other predictors, where we have readings in advance on a regular basis across elections. There are prediction markets that provide prices on a regular basis for many elections in the US and elsewhere, for example. Various other predictors are available over time in different election years, and these can be analyzed individually or in combination, as in Pollyvote (Graefe et al. 2014). Second, the method also can be used with other sets of events for which we have regular readings of predictors in advance. The most obvious examples may be the various events for which prediction markets exist, such as those relating to international relations, economics, sports and culture (Wolfers and Zitzewitz 2004).[18] Another set of applications might be where repeated measures converge on a final outcome, such as how sports teams' league places line up with their final position over the course of a season. Of course, actually applying the method requires a sufficient number of historical outcomes and observations over time. With the necessary data in hand, the general timeline approach – and much of what we have learned here – can be applied fairly directly. Indeed, Pathak et al. (2015) already have provided an initial foray by predicting Oscar winners from the flow of betting odds

[17] Also note that this approach allows the possibility of estimating fixed effects, which are of real consequence to comparative analysis (Jennings and Wlezien 2016).

[18] Here the precise application of the method will depend on whether the outcome variable in question is continuous (like vote share) or binary.

at different points in time leading up to the Oscars ceremony. What a more widespread application to other types of events would reveal remains to be seen.

REFERENCES

AAPOR. 2009. *An Evaluation of the Methodology of the 2008 Pre-Election Primary Polls: Prepared by the American Association for Public Opinion Research Ad Hoc Committee on the 2008 Presidential Primary Polling*. Lenexa, KS: American Association for Public Opinion Research.

Armstrong, J. Scott, Kesten C. Green, and Andreas Graefe. 2015. "Golden Rule of Forecasting: Be Conservative." *Journal of Business Research* 68(8):1717-1731.

Campbell, James E. 2008. *The American Campaign: US Presidential Campaigns and the National Vote*, 2nd edition. College Station, TX: Texas A&M University Press.

Erikson, Robert S. and Christopher Wlezien. 2012. *The Timeline of Presidential Elections: How Campaigns do (and do not) Matter*. Chicago: University of Chicago Press.

Graefe, Andreas. 2015. "German Election Forecasting: Comparing and Combining Methods for 2013." *German Politics* 24(2): 195-204.

Graefe, Andreas, Scott Armstrong, Randall Jones, and Alfred Cuzan. 2014. "Combining Forecasts: An Application to Elections." *International Journal of Forecasting* 30(1):43-54.

Honaker, James and Gary King. 2010. "What to do About Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(3): 561-581.

Jennings, Will and Christopher Wlezien. 2016. "The Timeline of Elections: A Comparative Perspective." *American Journal of Political Science* 60(1): 219-233.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1): 49-69.

Lall, Ranjit. 2016. “How Multiple Imputation Makes a Difference.” *Political Analysis*, first published online August 22, 2016 doi:10.1093/pan/mpw020

Lewis-Beck, Michael and Mary Stegmaier. 2014. “US Presidential Election Forecasting.” *PS: Political Science and Politics* 27(2):284-288.

Lewis-Beck, Michael and Charles Tien. 2016. “Election Forecasting: The Long View.” *Oxford Handbooks Online*. Oxford: Oxford University Press.

Pathak, Deepak, David Rothschild, and Miroslav Dudik. 2015. “A Comparison of Forecasting Methods: Fundamentals, Polling, Prediction Markets, and Experts.” *Journal of Prediction Markets* 9(2):1-31.

Pickup, Mark, and Richard Johnston. 2007. “Campaign Trial Heats as Electoral Information: Evidence from the 2004 and 2006 Canadian Federal Elections.” *Electoral Studies* 26(2): 460-476.

Rothschild, David. 2015. “Combining Forecasts for Elections: Accurate, Relevant, and Timely.” *International Journal of Forecasting* 31(3):952-964.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Silver, Nate. 2014. “Here’s Proof Some Pollsters Are Putting a Thumb on the Scale.” *FiveThirtyEight, November 14, 2014*. http://fivethirtyeight.com/features/heres-proof-some-pollsters-are-putting-a-thumb-on-the-scale/

Sturgis, Patrick, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, Jouni Kuha, Ben Lauderdale, and Patten Smith. 2016. *Report of the Inquiry into the 2015 British general election opinion polls*. London: Market Research Society/British Polling Council.

Wlezien, Christopher and Robert S. Erikson. 2001. "Campaign Effects in Theory and Practice." *American Politics Research* 29(5): 419-437.

Wlezien, Christopher and Robert S. Erikson. 2002. "The Timeline of Presidential Election Campaigns." *Journal of Politics* 64(4): 969-993.

Wlezien, Christopher, Will Jennings, Stephen Fisher, Robert Ford, and Mark Pickup. 2013. "Polls and the Vote in Britain." *Political Studies* 61(S1): 66-91.

Wolfers, Justin, and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives* 18(2):107-126.

APPENDIX

Table A1. An Analysis of Poll Variance, the Final 200 days of the Election Cycle

Variable			
$Party_{ij}$	578.3	-	376.8
$Country_{j}$	-	738.2	-
Year	-	-	38.9
Adjusted *R*-squared	0.86	0.44	0.90
RMSE	6.45	12.84	5.47

Number of parties = 255; number of countries = 45; number of elections = 249; number of party-election cases = 1,120; total number of observations = 23,760

Note: The estimates in the tables are F statistics.

Table A2. Correlations of Timelines Estimated Using Different Methods

	Raw data			
	Betas	**R^2**	**RMSE**	**MAE**
Betas				
R^2	0.8052			
	(0.0000)			
RMSE	-0.7791	-0.9402		
	(0.0000)	(0.0000)		
MAE	-0.8140	-0.8835	0.9386	
	(0.0000)	(0.0000)	(0.0000)	

	Interpolated data			
	Betas	**R^2**	**RMSE**	**MAE**
Betas				
R^2	0.9816			
	(0.0000)			
RMSE	-0.9895	-0.9981		
	(0.0000)	(0.0000)		
MAE	-0.9867	-0.9953	0.9969	
	(0.0000)	(0.0000)	(0.0000)	

	Interpolated data, bootstrapped			
	Betas	**R^2**	**RMSE**	**MAE**
Betas				
R^2	0.9816			
	(0.0000)			
RMSE	-0.9895	-0.9981		
	(0.0000)	(0.0000)		
MAE	-0.9867	-0.9953	0.9969	
	(0.0000)	(0.0000)	(0.0000)	

	Multiple imputation, bootstrapped			
	Betas	**R^2**	**RMSE**	**MAE**
Betas				
R^2	0.9819			
	(0.0000)			
RMSE	-0.9880	-0.9989		
	(0.0000)	(0.0000)		
MAE	-0.9908	-0.9941	0.9966	
	(0.0000)	(0.0000)	(0.0000)	

Note: *N* = 200; the numbers in parentheses are p-values.